DD2434 Machine Learning, Advanced Course

Assignment 1

Jiang, Sifan
sifanj@kth.se

December 8, 2019

1 The Prior $p(\mathbf{X})$, $p(\mathbf{W})$, $p(f)$

1.1 Theory

> **Question 1:** *Explain why Gaussian form of the likelihood is a sensible choice and what assumptions we make by this choice. What assumptions do we make about the data by choosing a spherical covariance matrix for the likelihood?*

For non-informative likelihoods, choosing Gaussian distributions is sensible because according to the *central limit theorem*, the sum of a set of random variables has a distribution that becomes increasingly Gaussian as the number of terms in the sum increases [1].

By choosing a spherical covariance matrix, we are assuming all the features $(x_1, x_2, \cdots, x_q)$ of the data are unrelated and independent. Also, the features have a same variance, which means, in the covariance matrix, all the diagonal elements are same while all non-diagonal elements are zero. In another word, the covariance $\Sigma = \sigma^2\mathbf{I}$.

> **Question 2:** *If we do **not** assume that the data points are independent how would the likelihood look then? Remember that $\mathbf{T} = [\mathbf{t}_1, \cdots, \mathbf{t}_N]$*

If we assume that the data points are independent, the likelihood would be

$$p(\mathbf{T}|f, \mathbf{X}) = \prod_{i=1}^{N} p(\mathbf{t}_i|f, \mathbf{x}_i). \tag{1}$$

However, if the data points are not independent, the likelihood of the data given the inputs and the mapping would be

$$\begin{aligned} p(\mathbf{T}|f, \mathbf{X}) &= \prod_{i=1}^{N} p(\mathbf{t}_i|f, \mathbf{x}_i, \mathbf{t}_1, \cdots, \mathbf{t}_{i-1}, \mathbf{t}_{i+1}, \cdots, \mathbf{t}_N) \\ &= \prod_{i=1}^{N} p(\mathbf{t}_i|f, \mathbf{x}_i, \mathbf{t}_1, \{\mathbf{T} \setminus \mathbf{t}_i\}) \end{aligned} \tag{2}$$

1.1.1 Linear Regression

> **Question 3:** *Please, complete the right-hand side of the expression in (6).*

Having the mapping:

$$\mathbf{t}_i = \mathbf{W}\mathbf{x}_i + \boldsymbol{\epsilon}, \tag{3}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, the likelihood of the data would be

$$p(\mathbf{T}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{t}_i|\mathbf{W}\mathbf{x}_i, \sigma^2\mathbf{I}). \tag{4}$$

Question 4: *The prior over each row of* $\mathbf{W}$ *in Eq.8 is a spherical Gaussian:* $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}_0, \tau^2\mathbf{I})$*. This means that the preferred model is encoded in terms of* L_2 *distance in the parameter space.*

- *What would be the effect of encoding the preferred model with* L_1 *norm (for model parameters)?*
- *Discuss how these two types of priors affect the posterior from the regularization perspective. Write down the penalization term, i.e. the negative log-prior, and illustrate for a two-dimensional problem (in the two-dimensional parameter space).*

- The prior over each row of $\mathbf{W}$ is a spherical Gaussian:

$$\begin{aligned} p(\mathbf{w}) =& \mathcal{N}\left(\mathbf{w}|\mathbf{w}_0, \tau^2\mathbf{I}\right) \\ =& \frac{1}{(2\pi)^{q/2}} \frac{1}{|\tau^2\mathbf{I}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^T(\tau^2\mathbf{I})^{-1}(\mathbf{w}-\mathbf{w}_0)\right], \end{aligned} \tag{5}$$

  so that the preferred model is encoded in terms of L_2 norm, since the quadratic function is used. If L_1 norm is used for encoding the preferred model, Manhattan distance from $\mathbf{w}$ to $\mathbf{w}_0$ would be calculated, $\sum_{i=1}^{q}\left|w^{(i)} - w_0^{(i)}\right|$, to encode the preferred model.

- The prior probability works as a regularization term in the exponent of the exponential part of the posterior. If both likelihood and prior are Gaussian, the posterior distribution would be

$$\begin{aligned} p(\mathbf{W}|\mathbf{X}, \mathbf{T}) =& \frac{1}{Z} p(\mathbf{T}|\mathbf{X}, \mathbf{W}) p(\mathbf{W}) \\ \propto& \exp\left[\sum_{i=1}^{N}\left(-\frac{1}{2}\left(\mathbf{t}_i - \mathbf{W}\mathbf{x}_i\right)^T\left(\sigma^2\mathbf{I}\right)\left(\mathbf{t}_i - \mathbf{W}\mathbf{x}_i\right)\right)\right. \\ & \left. -\frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^T(\tau^2\mathbf{I})^{-1}(\mathbf{w}-\mathbf{w}_0)\right]. \end{aligned} \tag{6}$$

  Then, to maximize the posterior distribution, we can minimize the negative log-posterior, which is

$$C + \frac{1}{2}\sum_{i=1}^{N}\left(\mathbf{t}_i - \mathbf{W}\mathbf{x}_i\right)^T\left(\sigma^2\mathbf{I}\right)\left(\mathbf{t}_i - \mathbf{W}\mathbf{x}_i\right) + \frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^T(\tau^2\mathbf{I})^{-1}(\mathbf{w}-\mathbf{w}_0), \tag{7}$$

  where C is a constant. Notice that

$$\frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^T(\tau^2\mathbf{I})^{-1}(\mathbf{w}-\mathbf{w}_0) \tag{8}$$

is the negative log-prior, which encoding the preferred model with L_2 norm, and is the penalization term, which is like the weight decay.

If the preferred model is encoded in terms of L_1 term. The penalization term, or negative log-prior, would be look like the lasso:

$$\frac{1}{2}\sum_{i=1}^{q}\frac{1}{\tau^2}\left|\mathbf{w}^{(i)}-\mathbf{w}_0^{(i)}\right|, \tag{9}$$

where $\mathbf{w}^{(i)}$ means the i-th element in $\mathbf{w}$, rows of $\mathbf{W}$.

A two-dimensional problem, meaning each $\mathbf{w}$ has only two elements: w_1 and w_2, and assuming $\mathbf{w}_0$ is the origin of coordinates, is illustrated in fig 1.

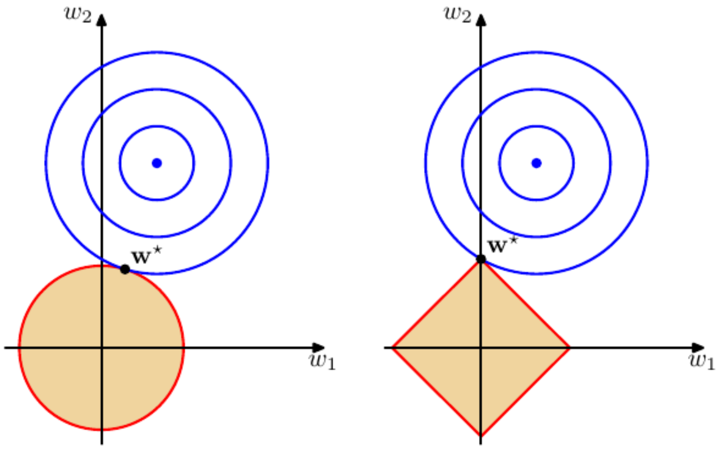


Figure 1: A two-dimensional problem. Left: L_2 norm. Right: L_1 norm. [1]

Question 5: *Assuming conditional independence of the target variables in* $\mathbf{t}$, *derive the posterior over the parameters. Please, do these calculations by hand as it is very good practice. To pass the assignment you only need to outline the calculation and highlight the important steps. In summary, please complete the following tasks*

- *Derive the posterior over the parameters and explain the final form in terms of the mean and covariance.*
- *How does the posterior form relate to the least square estimator of* $\mathbf{W}$ *(equivalent to the maximum likelihood approach) for this linear regression problem?*
- *How does the constant* Z *(eq 10) affect the solution? Are we interested in it?*

- The posterior is given by

$$p(\mathbf{W}|\mathbf{X},\mathbf{T}) = \frac{1}{Z}p(\mathbf{T}|\mathbf{X},\mathbf{W})p(\mathbf{W}), \tag{10}$$

where the likelihood $p(\mathbf{T}|\mathbf{X},\mathbf{W})$ is given in eq 4 such that

$$\begin{aligned} p(\mathbf{T}|\mathbf{X},\mathbf{W}) =& \prod_{i=1}^{N}\mathcal{N}(\mathbf{t}_i|\mathbf{W}\mathbf{x}_i,\sigma^2\mathbf{I}) \\ =& \mathcal{N}\left(\mathbf{T}|\mathbf{W}\mathbf{X},\sigma^2\mathbf{I}\right). \end{aligned} \tag{11}$$

The prior is given as

$$\begin{aligned} p(\mathbf{W}) =& \mathcal{MN}(\mathbf{W}_0, \mathbf{I}, \tau^2\mathbf{I}) \\ =& \mathcal{N}\left(\mathbf{W}|\mathbf{W}_0, \tau^2\mathbf{I}\right). \end{aligned} \tag{12}$$

Then, the posterior would be

$$\begin{aligned} p(\mathbf{W}|\mathbf{X}, \mathbf{T}) =& \frac{1}{Z}\mathcal{N}\left(\mathbf{T}|\mathbf{W}\mathbf{X}, \sigma^2\mathbf{I}\right)\mathcal{N}\left(\mathbf{W}|\mathbf{W}_0, \tau^2\mathbf{I}\right) \\ \propto& \exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{T}-\mathbf{W}\mathbf{X}\right)^T\left(\mathbf{T}-\mathbf{W}\mathbf{X}\right) - \frac{1}{2\tau^2}\left(\mathbf{W}-\mathbf{W}_0\right)^T\left(\mathbf{W}-\mathbf{W}_0\right)\right]. \end{aligned} \tag{13}$$

To find the mean and covariance for the posterior, we first find the components in a multivariate Gaussian distribution, which takes the form:

$$\begin{aligned} \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \Sigma\right) \propto& \exp\left[-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)^T\left(\Sigma\right)^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right] \\ =& \exp\left[-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x} + \mathbf{x}^T\Sigma^{-1}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}\right]. \end{aligned} \tag{14}$$

Notice that the exponent of the general conditional multivariate Gaussian distribution has three parts: *quadratic*, *linear*, and *constant* parts. The method to find the mean and covariance for posterior is to form those three parts in the exponent of the distribution $p\left(\mathbf{W}|\mathbf{X}, \mathbf{T}\right)$.

$$\begin{aligned} & -\frac{1}{2\sigma^2}\left(\mathbf{T}-\mathbf{W}\mathbf{X}\right)^T\left(\mathbf{T}-\mathbf{W}\mathbf{X}\right) - \frac{1}{2\tau^2}\left(\mathbf{W}-\mathbf{W}_0\right)^T\left(\mathbf{W}-\mathbf{W}_0\right) \\ =& -\frac{1}{2\sigma^2}\mathbf{T}^T\mathbf{T} + \frac{1}{\sigma^2}\mathbf{T}^T\mathbf{W}\mathbf{X} - \frac{1}{2\sigma^2}\left(\mathbf{W}\mathbf{X}\right)^T\left(\mathbf{W}\mathbf{X}\right) \\ & -\frac{1}{2\tau^2}\mathbf{W}^T\mathbf{W} + \frac{1}{\tau^2}\mathbf{W}^T\mathbf{W}_0 - \frac{1}{2\tau^2}\left(\mathbf{W}_0\right)^T\left(\mathbf{W}_0\right) \\ =& -\frac{1}{2\sigma^2}\left(\mathbf{W}\mathbf{X}\right)^T\left(\mathbf{W}\mathbf{X}\right) - \frac{1}{2\tau^2}\mathbf{W}^T\mathbf{W} \\ & +\frac{1}{\sigma^2}\mathbf{T}^T\mathbf{W}\mathbf{X} + \frac{1}{\tau^2}\mathbf{W}^T\mathbf{W}_0 \\ & -\frac{1}{2\sigma^2}\mathbf{T}^T\mathbf{T} - \frac{1}{2\tau^2}\left(\mathbf{W}_0\right)^T\left(\mathbf{W}_0\right) \end{aligned} \tag{15}$$

where the *quadratic* term can be written as

$$-\frac{1}{2}\mathbf{W}^T\left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\tau^2}\mathbf{I}\right)\mathbf{W}, \tag{16}$$

so that the posterior covariance would be

$$\Sigma = \left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\tau^2}\mathbf{I}\right)^{-1}. \tag{17}$$

The *linear* term can be written as

$$\mathbf{W}^T\Sigma^{-1}\boldsymbol{\mu}, \tag{18}$$

so that the posterior mean would be

$$\boldsymbol{\mu} = \left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\tau^2}\mathbf{I}\right)^{-1}\left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\tau^2}\mathbf{W}_0\right). \tag{19}$$

With the Σ and $\boldsymbol{\mu}$, the posterior distribution would be $\mathcal{N}\left(\mathbf{W}|\boldsymbol{\mu},\Sigma\right)$.

- The posterior form is the least square estimator of $\mathbf{W}$ added with a regularization term, which is the prior distribution.
- Z is the normalizing constant that makes sure the integral of the distribution equals to 1. We are not interested in it when maximizing the posterior since the scaling won't affect the result that maximize the distribution. Also, since Gaussian distributions are conjugate to each other, so that we would always get a normalized distribution explained by the mean and the covariance.

1.1.2 Non-parametric Regression

Question 6: *Explain what this prior does. Motivate the choice of this prior and use images to show your reasoning. Clue: use the marginal distribution to explain the prior*

Since we have

$$\begin{aligned}\mathbf{t}_i =& f(\mathbf{x}_i) + \boldsymbol{\epsilon} \\ =& f_i + \boldsymbol{\epsilon},\end{aligned} \tag{20}$$

then, if consider noise processes that have a Gaussian distribution, the conditional distribution on $\mathbf{t}_i$ would be

$$p(\mathbf{t}_i|f_i) = \mathcal{N}\left(\mathbf{t}_i|f_i, \beta^{-1}\mathbf{I}_D\right), \tag{21}$$

where β is a hyperparameter representing the precision of the noise and $\mathbf{I}_D$ is the $D \times D$ unit matrix (notice that $\mathbf{t}_i \in \mathbb{R}^D$). According to Bishop [1], since we not have any prior knowledge about the mean of $\mathbf{t}$ and so by symmetry we take it to be zero, which is equivalent to choosing the mean of the prior over f_i to be zero. The set of values f is $(f(\mathbf{x}_1), \cdots, f(\mathbf{x}_N))$, so the marginal distribution, which is the prior, is given by a Gaussian whose mean is zero and whose covariance is defined by a Gram matrix $\mathbf{K} = k(\mathbf{X},\mathbf{X})$ so that

$$p\left(f|\mathbf{X},\boldsymbol{\theta}\right) = \mathcal{N}\left(f|\mathbf{0}, k(\mathbf{X},\mathbf{X})\right), \tag{22}$$

where $k(\cdot,\cdot)$ is the covariance and $\boldsymbol{\theta}$ is its parameters as described. The kernel function k express the property that, for input point sets $\mathbf{x}_n$ and $\mathbf{x}_m$ that are similar, the corresponding values $f(\mathbf{x}_n)$ and $f(\mathbf{x}_m)$ will be more strongly correlated than for dissimilar points [1]. The correlation is controlled by the hyperparameter $\boldsymbol{\theta}$.

For example, we can choose RBF kernel in $\mathcal{GP}$ regression:

$$k\left(\mathbf{x}_n,\mathbf{x}_m\right) = \theta_0 \exp\frac{\theta_1}{2}\|\mathbf{x}_n - \mathbf{x}_m\|^2 + \theta_2 + \theta_3\mathbf{x}_n^T\mathbf{x}_m \tag{23}$$

Then, the effect of the hyperparameters in illustrated in fig 2.

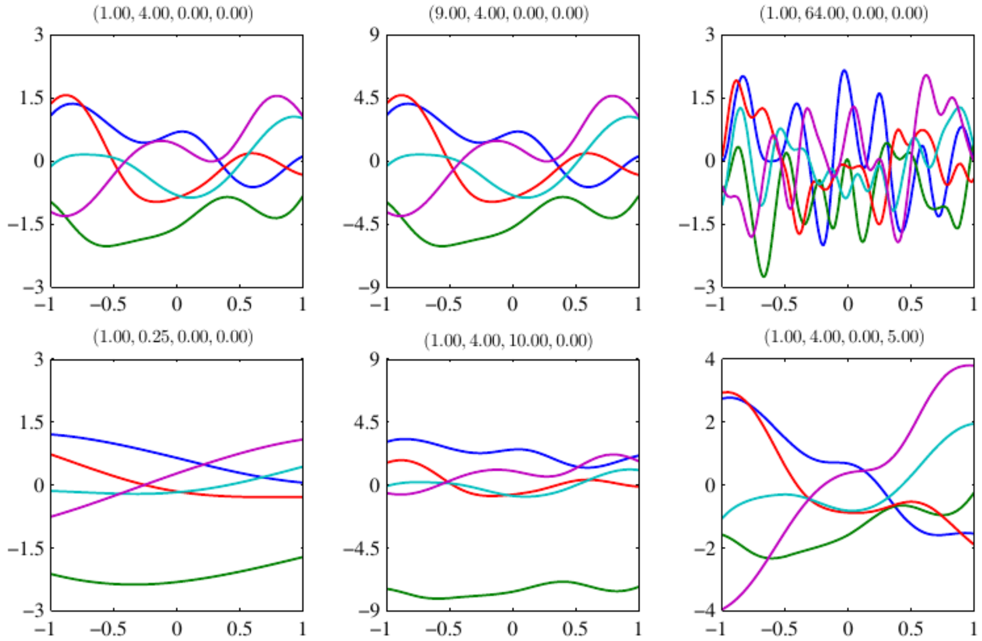


Figure 2: Samples from a Gaussian process prior [1].

Question 7: *Formulate the joint likelihood of the full model defined above,*

$$p(\mathbf{T}, \mathbf{X}, f, \boldsymbol{\theta})$$

and draw a simple graphical model reflecting the assumptions that you have made.

According to eq 20 and 22, we can conclude that the target $\mathbf{T}$ directly depends on f and f is directly depends on $\mathbf{X}$ and $\boldsymbol{\theta}$. And it is natural and reasonable to assume that $\mathbf{X}$ and $\boldsymbol{\theta}$ are independent, so the joint likelihood of the full model can be written as

$$p(\mathbf{T}, \mathbf{X}, f, \boldsymbol{\theta}) = p(\mathbf{T}|f)p(f|\mathbf{X}, \boldsymbol{\theta})p(\mathbf{X})p(\boldsymbol{\theta}). \tag{24}$$

The graphic model reflecting this assumption is illustrated in fig 3.

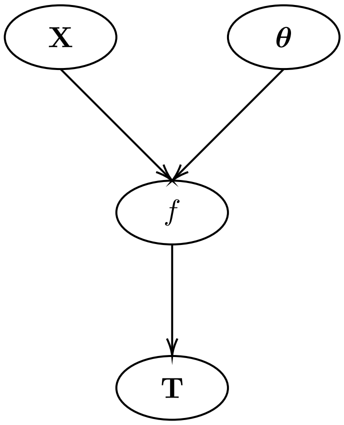


Figure 3: Graphic model for question 7.

Question 8: *Complete the marginalisation formula in Eq.12 (general form) and discuss the following:*

- *Explain how it connects the prior and the data.*
- *How does the uncertainty "filter" through the marginalisation?*
- *Why do we still condition on $\boldsymbol{\theta}$ after the marginalisation?*

In order to marginalize to obtain the distribution $p(\mathbf{T}|\mathbf{X}, \boldsymbol{\theta})$, which conditioned on the input values $\mathbf{x}_1, \cdots, \mathbf{x}_N$, we need to integrate over f, so that we get

$$p(\mathbf{T}|\mathbf{X}, \boldsymbol{\theta}) = \int p(\mathbf{T}|f)p(f|\mathbf{X}, \boldsymbol{\theta})df. \tag{25}$$

- $p(f|\mathbf{X}, \boldsymbol{\theta})$ is the prior, where the non-linear function f and all the input data $\mathbf{X}$ are considered. By marginalizing over f, the input data $\mathbf{X}$ in the prior and the target data $\mathbf{T}$ in the likelihood are connected. So, we can get the probability of the target data given the input data and the hyperparameter.
- The uncertainty of having the function to get the target data is implied in the likelihood $p(\mathbf{T}|f)$, and the uncertainty of having the input data and the hyperparameter to get the function is implied in the prior $p(f|\mathbf{X}, \boldsymbol{\theta})$. By marginalization, the uncertainties from two parts are added so that the uncertainty is "filtered" to give uncertainty of having the input data and the hyperparameter to get the target data. Also, marginalization is like an ensemble learning that averages out uncertainty, which is like "filtering".
- Since the function f is generated according to $\mathbf{X}$ and $\boldsymbol{\theta}$, after marginalizing on f, the "effect" of $\boldsymbol{\theta}$ still exists. To deal with $\boldsymbol{\theta}$, Bayesian approach can be applied to identify the prior and posterior and average it out.

1.2 Practical

1.2.1 Linear Regression

Question 9:

1. *Set the prior distribution over* $\mathbf{W}$ *and visualise it.*
2. *Pick a single data point* (x, t) *and visualise the posterior distribution over* $\mathbf{W}$*.*
3. *Draw* 5 *samples from the posterior and plot the resulting functions.*
4. *Repeat 2 - 3 by adding additional data points up to* 7*.*
5. *Given the plots explain the effect of adding more data on the posterior as well as the functions. How would you interpret this effect?*
6. *Finally, test the exercise for different values of* σ^2*, e.g.* 0.1*,* 0.4 *and* 0.8*. How does your model account for data with varying noise levels? What is the effect on the posterior?*

1. Since $\mathbf{W}$ is non-informative, we can simply set is the prior distribution over it as a multivariate normal distribution with zero mean and unit covariance, which is visualized in fig 4.

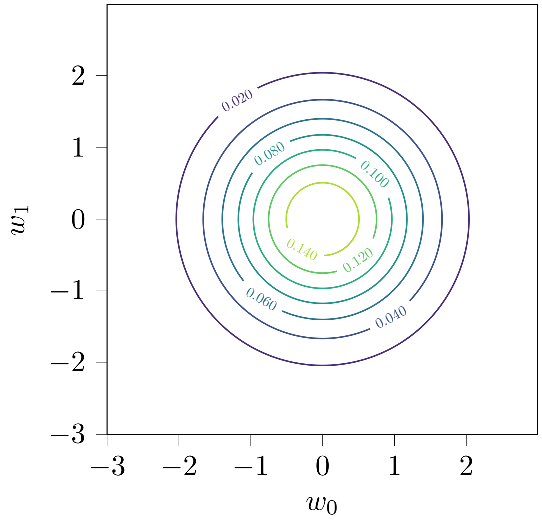
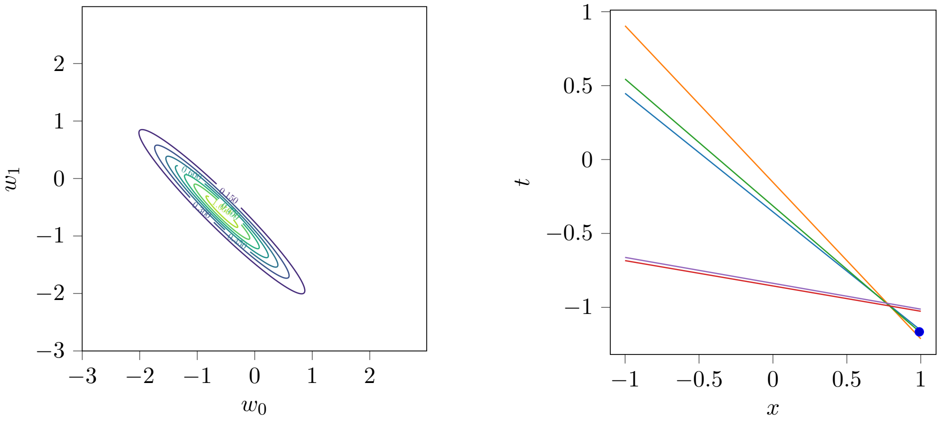


Figure 4: The prior distribution over $\mathbf{W}$.

2. The random seed used in this question and questions is 1000 (for both *random* and *numpy.random*). After picking a single data point (x_1, t_1), the posterior can be obtained through eq 10 and applied by the eq 17 and 19. The posterior distribution over $\mathbf{W}$ is illustrated in fig 5a.



(a) The posterior distribution over $\mathbf{W}$.

(b) The resulting functions.

Figure 5: Posterior distribution and resulting functions with 2 data points.

3. After drawing 5 samples from the posterior, the resulting functions is illustrated in fig 5b. The blue point (also in the following questions) is the observed data point.

4. The results are shown from fig 6 to 11.

5. By adding more data, the posterior distribution over $\mathbf{W}$ would have mean which much closer to the correct value, $[0.5, -1.5]$, and have covariance which becomes much smaller.

   For the functions, after the data points are added, the average slope approaches to 0.5 and the average intercept approaches to -1.5 at $x = 0$. Also, the difference between the slopes and the intercepts of the five functions decreases by adding more data.

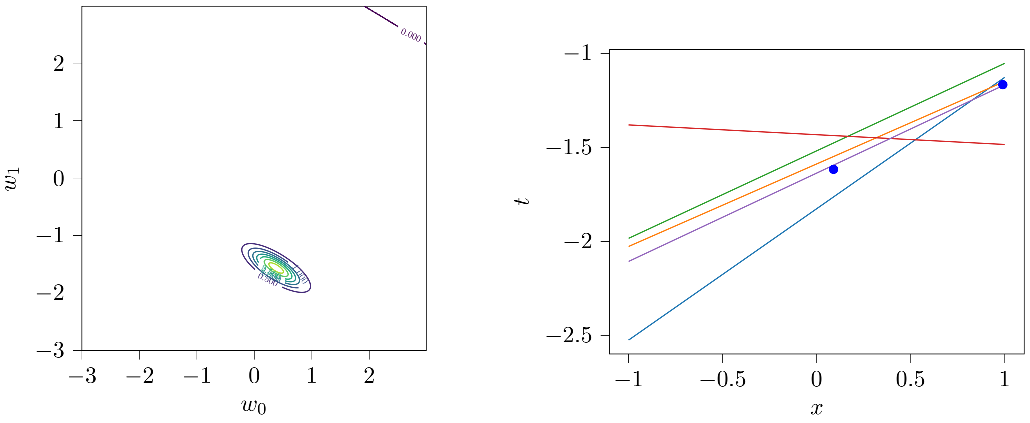


Figure 6: Posterior distribution and resulting functions with 2 data points.

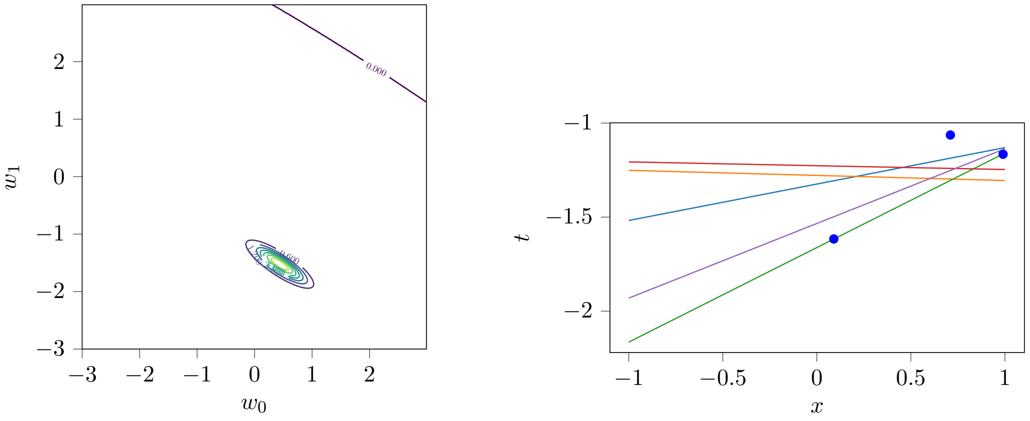


Figure 7: Posterior distribution and resulting functions with 3 data points.

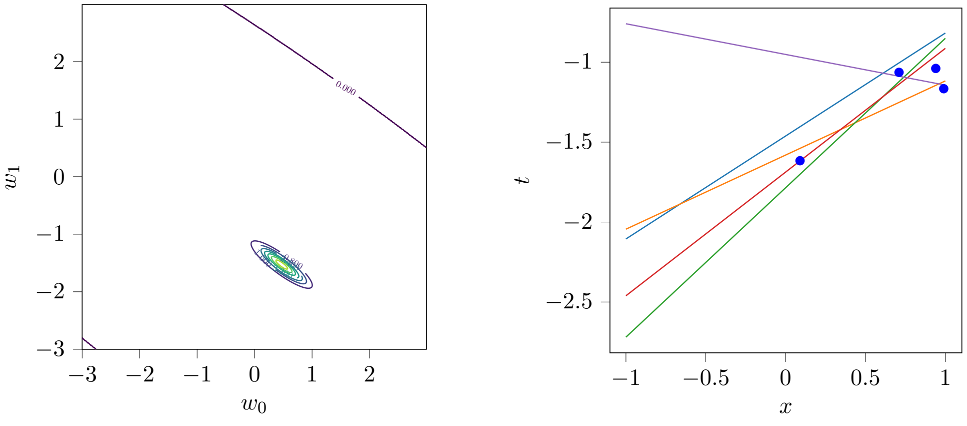


Figure 8: Posterior distribution and resulting functions with 4 data points.

Since the data point is generated from the linear function with noise, which is a normal distribution with zero mean, the data generated is following the normal distribution where the mean is the linear function and the variance is the variance from the noise. This means,

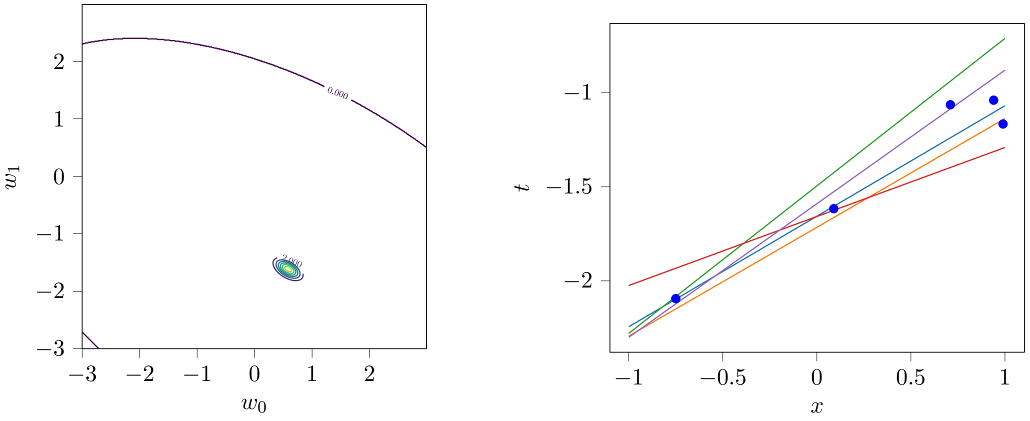


Figure 9: Posterior distribution and resulting functions with 5 data points.

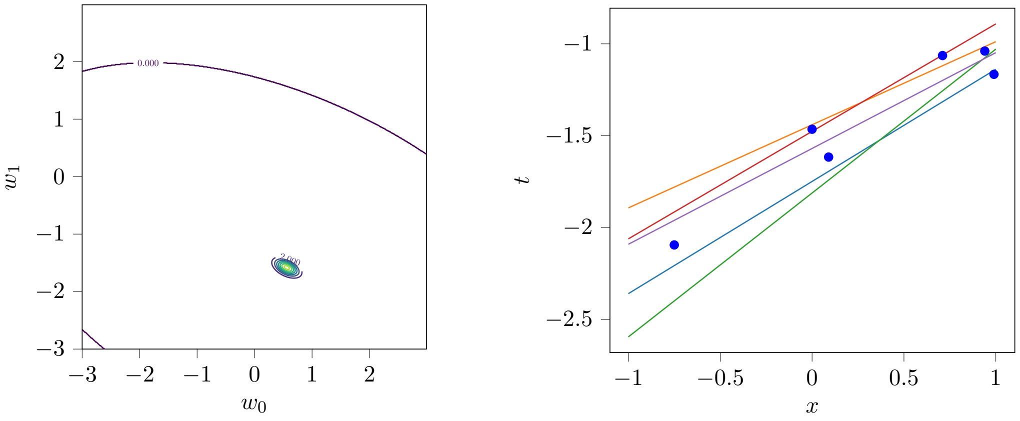


Figure 10: Posterior distribution and resulting functions with 6 data points.

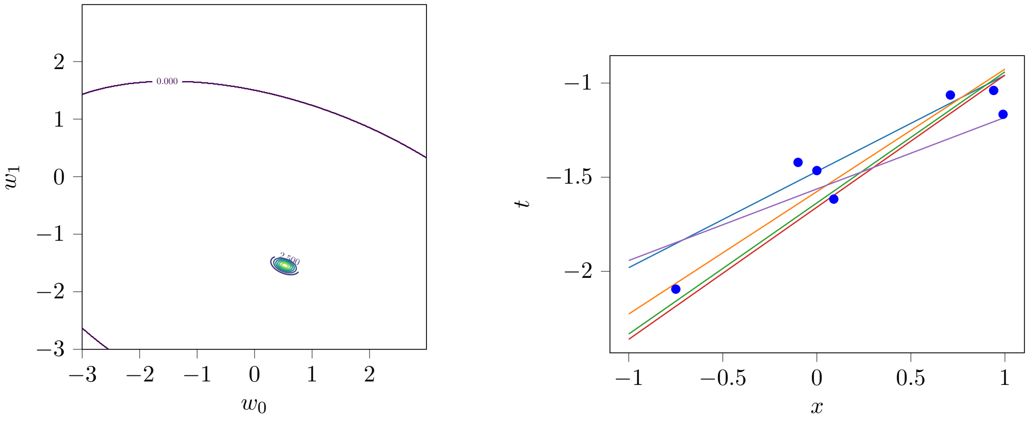
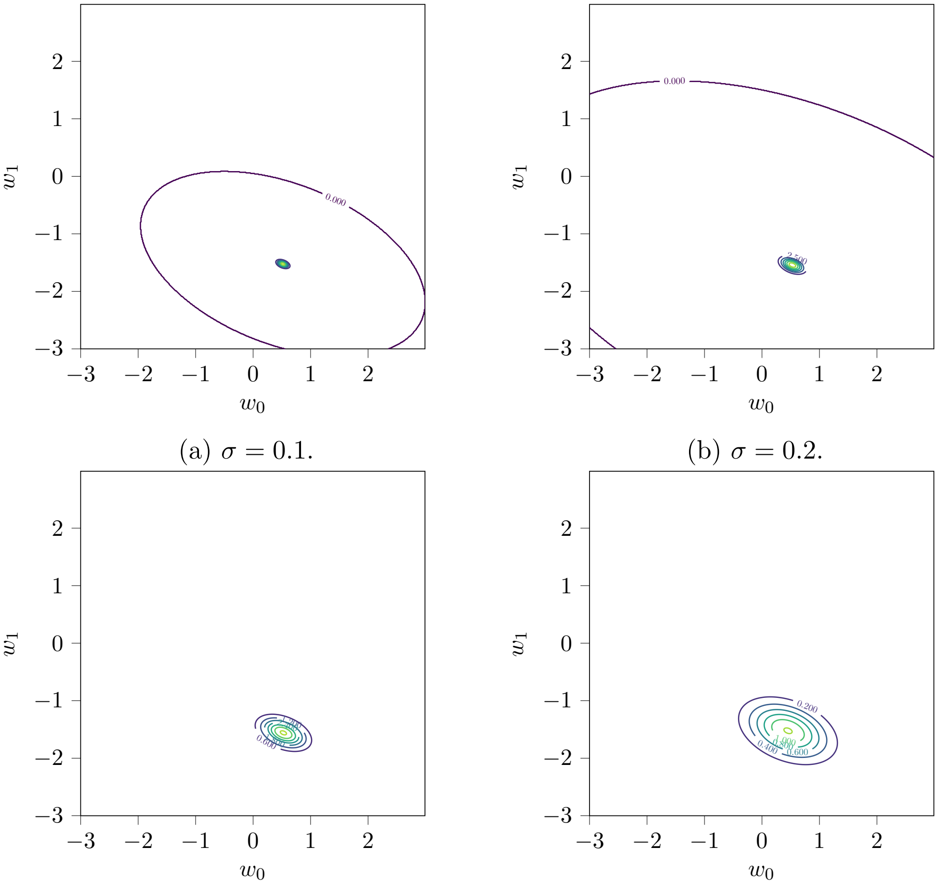


Figure 11: Posterior distribution and resulting functions with 7 data points.

the more data we add, the data points close to linear function would have bigger probability. Then using such data, the posterior like over $\mathbf{W}$ would have mean that closer to the model that generates the data and with more data, the confidence increases. Then, since the functions is directed generated from the posterior distribution over **W**, the functions would have the feature of the distribution.

6. The posterior distribution over **W** after having 7 data points but with different values of σ in $0.1, 0.2, 0.4, 0.8$ is shown in fig 12. Also, the corresponding sampled functions are shown in fig 13. To make it reasonable to compare the effect of the values of σ, for each σ, the random function is reset after at the beginning. As you can see on fig 13, the patterns of the observed data are the same, the difference between the data points from different test is caused by the variance of the model, so that the data points with bigger value of σ are much sparser than the ones with smaller values.

   With less precise data points, the corresponding posterior distribution over **W** would have larger covariance. However, as you can see on fig 12, since the data points are generated through same random seed or pattern, the mean of each models are similar. Then, as discussed in the previous problem, the functions sampled from the posterior distributions would present the features of the distributions in the way how big is the slope and intercept and how the functions are different to each other.



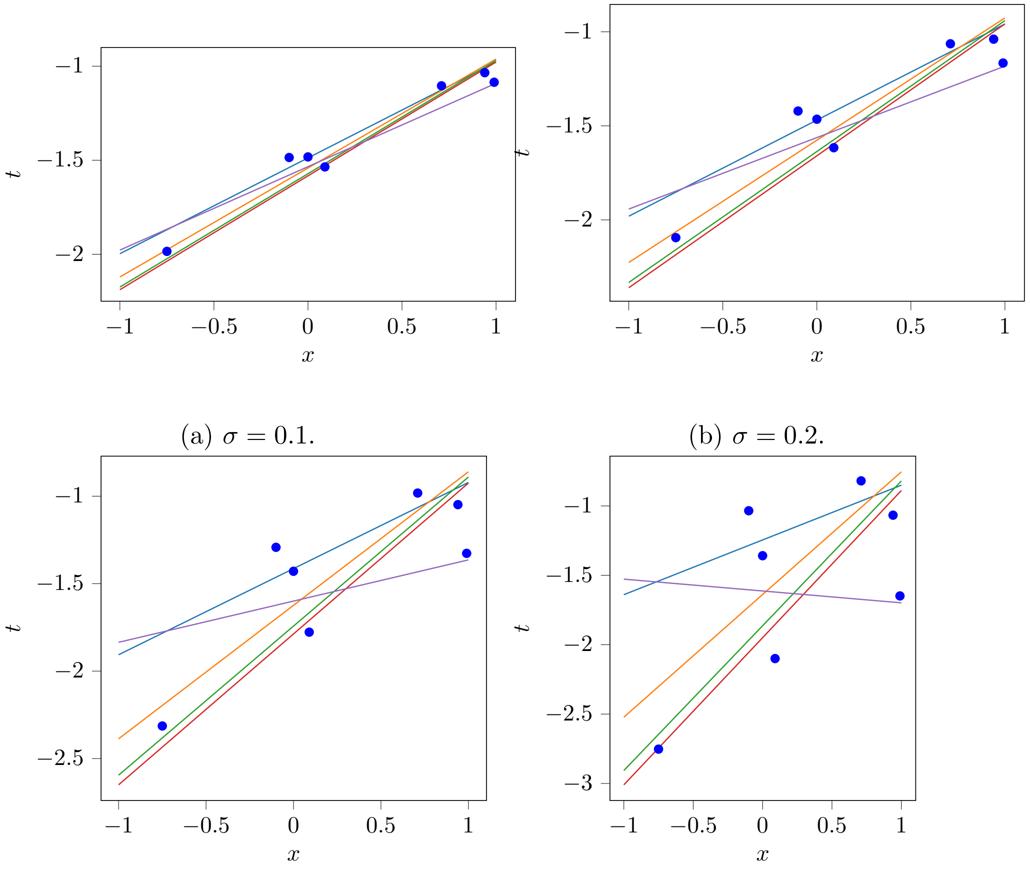
(a) $\sigma = 0.1$.

(b) $\sigma = 0.2$.

(c) $\sigma = 0.4$.

(d) $\sigma = 0.8$.

Figure 12: Posterior distribution over **W** with 7 data points with different values of σ.



(a) $\sigma = 0.1$.

(b) $\sigma = 0.2$.

(c) $\sigma = 0.4$.

(d) $\sigma = 0.8$.

Figure 13: The resulting functions with 7 data points with different values of σ.

1.2.2 Non-parametric Regression

Question 10:

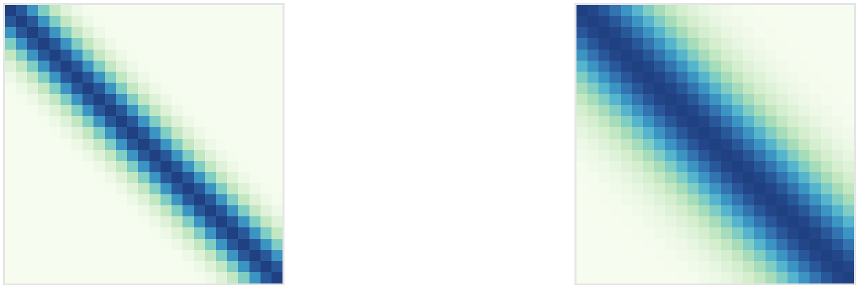
1. *Create a $\mathcal{GP}$-prior with a squared exponential covariance function.*
2. *For each of 4 different length scales, please draw 10 samples from this prior and visualise them. Explain the observed consequences of altering the length-scale of the covariance function.*

1. The $\mathcal{GP}$-prior with a squared exponential covariance function is:

$$p\left(\mathbf{f}|\mathbf{X},\boldsymbol{\theta}\right) = \mathcal{N}\left(\mathbf{f}|\mathbf{0},\mathbf{K}\right), \tag{26}$$

   where $\mathbf{K}$ is the covariance matrix given by the kernel function. The squared exponential kernel with $\sigma_f = 1.0$ and $l = 1.0$ or 2.0 is illustrated in fig 14.

2. The result is shown in fig 15. The length scales l have the effect on the smoothness of the prior functions: as l gets bigger, the functions become much smoother. From the equation of the squared exponential covariance function, we can know that when the length scales approaches to positive infinity, the exponent approaches to zero and the kernel function approaches to its maximal value, σ_f^2, which means for all the $(\mathbf{x}_i, \mathbf{x}_j)$, the kernel would give the same value.

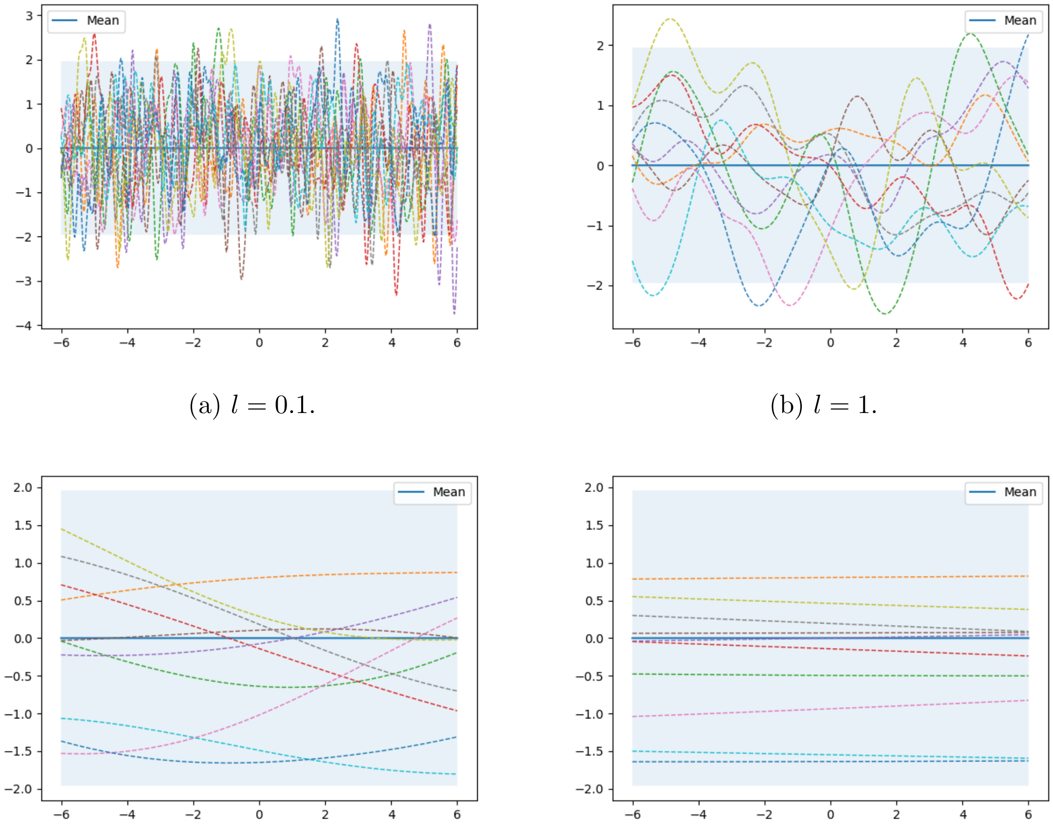


(a) $l = 1.0$ and $\sigma_f = 1.0$.

(b) $l = 2.0$ and $\sigma_f = 1.0$.

Figure 14: RBF kernel with different length scales.

Is this question, the value of σ_f is assume to be 1.0, then a larger length scale would implies more correlation between the dimensions of $\mathbf{X}$, which can be seen from fig 14.



(a) $l = 0.1$.

(b) $l = 1$.

(c) $l = 10$.

(d) $l = 100$.

Figure 15: Samples from the $\mathcal{GP}$-prior with different length scales.

Question 11:

1. *What is the posterior before we observe any data?*
2. *Compute the predictive posterior distribution of the model.*
3. *Sample from this posterior with points both close to and far away from the observed data. Explain the observed effects.*
4. *Plot the data, the predictive mean and the predictive variance of the posterior from the data.*
5. *Compare the samples of the posterior with the ones from the prior. Is the observed behavior desirable?*
6. *What would happen if you added a diagonal covariance matrix to the squared exponential?*

1. Before we observed any data, the posterior would be the same with the $\mathcal{GP}$-prior, with is

$$\mathcal{N}\left(\mathbf{f}|\mathbf{0}, \mathbf{K}\right), \tag{27}$$

where $\mathbf{K}$ is the Gram matrix given by the covariance function.

2. A $\mathcal{GP}$-prior $p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta})$ can be converted into a $\mathcal{GP}$-posterior $p(\mathbf{f}|\mathbf{T}, \mathbf{X}, \boldsymbol{\theta})$ after observing data. Then the posterior distribution can be used to predict unobserved data. The predictive posterior distribution of the model is given by $p(t_{N+1}|\mathbf{t}_N)$, where t_{N+1} is the target to be predicted for a new input $\mathbf{x}_{N+1}$ based on observed targets $\mathbf{t}_N$. Note that this distribution is conditioned also on the variables $\mathbf{x}_1, \cdots, \mathbf{x}_N$ and $\mathbf{x}_{N+1}$ [1]. To find this distribution, we can start with the joint distribution $p(\mathbf{t}_{N+1})$, where $\mathbf{t}_{N+1}$ denotes vector $(\mathbf{t}_1, \cdots, \mathbf{t}_N, \mathbf{t}_{N+1})$, which is

$$p(\mathbf{t}_{N+1}) = \mathcal{N}\left(\mathbf{t}_{N+1}|\mathbf{0}, \mathbf{C}_{N+1}\right), \tag{28}$$

where

$$\mathbf{C}_{N+1} = \begin{pmatrix} \mathbf{C}_N & \mathbf{k} \\ \mathbf{k}^T & c \end{pmatrix}, \tag{29}$$

where $\mathbf{C}_N$ is the $N \times N$ covariance matrix in the $p(\mathbf{t}_N)$, so that $C(\mathbf{x}_n, \mathbf{x}_m) = k(\mathbf{x}_n, \mathbf{x}_m) + \beta^{-1}\delta_{nm}$, where $\beta^{-1} = 0.3$. The vector $\mathbf{k}$ has elements $k(\mathbf{x}_n, \mathbf{x}_{N+1})$ for $n = 1, \cdots, N$, and the scalar $c = k\left(\mathbf{x}_{N+1}, \mathbf{x}_{N+1}\right) + \beta^{-1}$. The conditional distribution $p(t_{N+1}|\mathbf{t})$ is a Gaussian distribution with mean and covariance given by

$$\begin{aligned} m(\mathbf{x}_{N+1}) =& \mathbf{k}^T\mathbf{C}_N^{-1}\mathbf{t} \qquad (30) \\ \sigma^2(\mathbf{x}_{N+1}) =& c - \mathbf{k}^T\mathbf{C}_N^{-1}\mathbf{k}. \qquad (31) \end{aligned}$$

3. After generating data (with noise $\beta^{-1} = 0.3^2$), the model is trained with the training data and then the predictive posterior distribution of the model is obtained by the method shown in the above question. Then three samples are generated from this distribution which is shown in fig 16, where the length scale used is 10. As shown on the plot, the samples are much closer to each other at the points close to the observed data, while diverge a lot at the points far away from the observed data. That's because at the points near the observed data, information can reduce the variance of the distribution, thus the predictive samples would

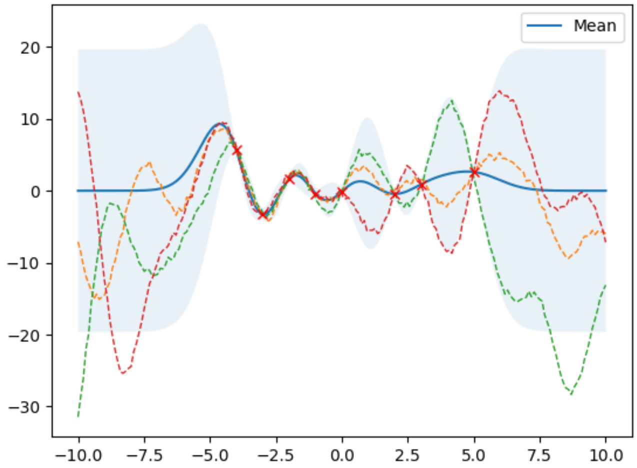


Figure 16: Samples from the predictive posterior distribution of the model with $l = 10$.

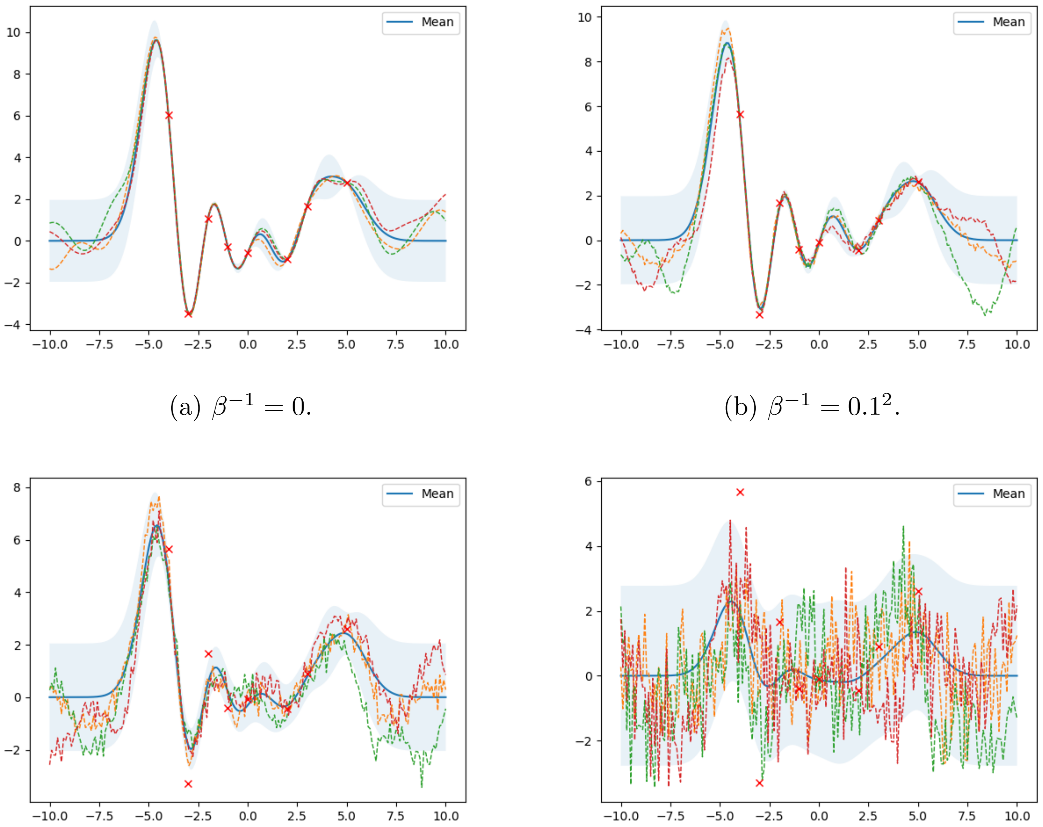
be close to the mean value. While at the points that are far away from the observed data, little or no information is obtained so that the variance is big and the corresponding samples would diverge from each other.

4. The data are plotted in the fig 16, where the blue solid curve is the mean, the light blue shows the variance (σ multiplied by a constant) of the posterior from the data. The red crosses are the observed data points.

5. The samples of the prior is shown in fig 15 and the samples of the posterior is shown in fig 16. The mean of the samples of the posterior distribution moves towards to the ground truth function at the points near to the observed data compared to the samples of the prior distribution. Also the variance shrinks where near those points while the expands where far away from those observed data points. Such behaviors are desirable because after the prior distribution is updated, informative points have much confidence on the predictive values while the non-informative points are not confident with the predictive results.

6. In the former problems, the diagonal covariance matrix is already be added to the predictive posterior distribution. so, to see the effect of adding the diagonal covariance matrix to the squared exponential, we can first see what would happen when there is no diagonal covariance matrix, then we increase the σ in the normal distribution of ϵ. The resulting plot is shown in fig 17.

   From the plot, we can see that when there's no covariance matrix, the variance at the observed points would also be zero and small near them and the mean of the distribution would exactly pass through the observed points. Then if a diagonal covariance matrix is added to the squared exponential, the variance at the observed points would expand, and the mean would deviate from the observed points. The bigger the covariance matrix is, the bigger the variance and the larger the deviation would be.



(a) $\beta^{-1} = 0$.

(b) $\beta^{-1} = 0.1^2$.

(c) $\beta^{-1} = 0.3^2$.

(d) $\beta^{-1} = 1^2$.

Figure 17: Samples from the posterior distribution with different diagonal matrices.

2 The Posterior $p(\mathbf{X}|\mathbf{Y})$

2.1 Theory

Question 12: *What type of "preference" for the latent variable* $\mathbf{X}$ *does this prior encode?*

Since the prior over the latent variables is a spherical Gaussian,

$$p(\mathbf{X}) = \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{32}$$

which encodes the "preference" for the latent variable $\mathbf{X}$, where the dimensions are not correlated or independent with each other, meaning only the element on the diagonal has value (which is 1) while others are zero. Also, since $\mathbf{X}$ and $\mathbf{W}$ have relation (in linear it's easy to get), having the preference on $\mathbf{X}$ also means having preference on $\mathbf{W}$.

Question 13: *Perform the marginalisation in Eq. 23 and write down the expression. As previously, it is recommended that you do this by hand. In the answer outline the calculations and highlight the important steps.*
Hint: *The marginal can be computed by integrating out* $\mathbf{X}$ *with the use of Gaussian algebra we exploited in the exercise derivations and, in particular, by completing the square. However, it is much easier to derive the mean and covariance, knowing that the marginal is Gaussian, from the linear equation of* $\mathbf{Y}(\mathbf{X})$*.*

To perform the marginalisation on

$$p(\mathbf{Y}|\mathbf{W}) = \int p(\mathbf{Y}|\mathbf{X}, \mathbf{W})p(\mathbf{X})d\mathbf{X}, \tag{33}$$

we start from the linear model for single data, which is

$$\mathbf{y}_i = \mathbf{W}\mathbf{x}_i + \boldsymbol{\epsilon}, \tag{34}$$

where $\mathbf{y}_i \in \mathbb{R}^D$, $\mathbf{x}_i \in \mathbb{R}^q$, and $\boldsymbol{\epsilon} \in \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Then the likelihood would be

$$p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{W}) = \mathcal{N}\left(\mathbf{y}_i|\mathbf{W}\mathbf{x}_i, \sigma^2\mathbf{I}\right). \tag{35}$$

Then the marginalisation would be obtained by integrating over the latent variables $\mathbf{X}$:

$$p(\mathbf{y}_i|\mathbf{W}) = \int p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{W})p(\mathbf{x}_i)d\mathbf{x}_i. \tag{36}$$

Since the marginal distribution is Gaussian, we can derive the mean and covariance (and since there's no $\mathbf{y}_i$ in the prior of $\mathbf{x}_i$, it is almost straight forward to get the mean):

$$\begin{aligned}
\mu_{\mathbf{y}_i|\mathbf{W}} =& \mu_{\mathbf{y}_i|\mathbf{x}_i,\mathbf{W}} \\
=& \mathbb{E}\left[\mathbf{W}\mathbf{x}_i + \boldsymbol{\epsilon}\right] \\
=& \mathbf{W}\mathbb{E}\left[\mathbf{x}_i\right] + \mathbb{E}[\boldsymbol{\epsilon}] \\
=& \mathbf{W}\mu_{\mathbf{x}_i} + \mu_{\boldsymbol{\epsilon}} \\
=& \mathbf{0},
\end{aligned} \tag{37}$$

and based on $\mathrm{Var}(A) = \mathbb{E}[(A - \mathbb{E}[A])^2]$, the covariance could be get:

$$\begin{aligned}
\Sigma_{\mathbf{y}_i|\mathbf{W}} =& \mathbb{E}\left[\left(\mathbf{y}_i|\mathbf{W} - \mathbb{E}\left[\mathbf{y}_i|\mathbf{W}\right]\right)\left(\mathbf{y}_i|\mathbf{W} - \mathbb{E}\left[\mathbf{y}_i|\mathbf{W}\right]\right)^T\right] \\
=& \mathbb{E}\left[\left(\mathbf{W}\mathbf{x}_i + \boldsymbol{\epsilon}\right)\left(\mathbf{W}\mathbf{x}_i + \boldsymbol{\epsilon}\right)^T\right] \\
=& \mathbb{E}\left[\mathbf{W}\mathbf{x}_i\mathbf{x}_i^T\mathbf{W}^T + 2\mathbf{W}\mathbf{x}_i\boldsymbol{\epsilon}^T + \boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right] \\
=& \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}.
\end{aligned} \tag{38}$$

Then the marginal distribution can be obtained by production of $\mathbf{y}_i$:

$$p\left(\mathbf{Y}|\mathbf{W}\right) = \prod_{i=1}^{N}\mathcal{N}\left(\mathbf{y}_i|\mu_{y_i|\mathbf{W}}, \Sigma_{\mathbf{y}_i|\mathbf{W}}\right). \tag{39}$$

2.1.1 Learning

Question 14: *Compare the three different estimation procedures above in log-space.*

1. *What are their distinctive features and how are they different when we observe more data?*
2. *Why are the two last expressions of Eq. 25 equal?*
3. *Explain why Type-II Maximum-Likelihood is a sensible approach to learn the model.*

1. Maximizing likelihood is equivalent to minimizing the log-likelihood, which is equal to:

$$\hat{\mathbf{W}} = \arg\min_{\mathbf{W}} \sum_{i=1}^{N} (\mathbf{y}_i - \mathbf{W}\mathbf{x}_i)^T (\mathbf{y}_i - \mathbf{W}\mathbf{x}_i) . \tag{40}$$

   With same idea, MAP is equivalent to:

$$\hat{\mathbf{W}} = \arg\min_{\mathbf{W}} \left[\frac{1}{\sigma^2} \sum_{i=1}^{N} (\mathbf{y}_i - \mathbf{W}\mathbf{x}_i)^T (\mathbf{y}_i - \mathbf{W}\mathbf{x}_i) + \frac{1}{\tau^2} \sum \mathbf{w}_i^2 \right] . \tag{41}$$

   And the Type-II Maximum-Likelihood is equivalent to:

$$\hat{\mathbf{W}} = \arg\min_{\mathbf{W}} \left[N \ln \left(|\mathbf{W}\mathbf{W}^T + \sigma^2 \mathbf{I}| \right) + \sum_{i=1}^{N} \mathbf{y}_i^T \left(\mathbf{W}\mathbf{W}^T + \sigma^2 \mathbf{I} \right)^{-1} \mathbf{y}_i \right] . \tag{42}$$

   Comparing those three different estimation procedures, we can tell that MLE is a special case of MAP when the regularized term is equal to zero. So, MLE would be easy to compute without doing any matrix inverse. MAP introduce the prior which would give a regularized term that can prevent the learning from overfitting. Type-II Maximum-Likelihood marginalized one latent variable, so during the estimation procedure, this latent variable won't be taken into consideration.

2. The two last expression of Eq. 25 are equal because the integration on the denominator marginalizes over **W**, meaning it won't affect the estimation procedure, so we can just remove it.

3. Type-II Maximum Likelihood is a sensible approach to learn the model because the marginal distribution is marginalized over **X** with is not contribute to the estimation procedure. Compared to MLE, it don't require to take **X** into the computation. Also, since **X** and **W** are related, marginalization over one parameter would prevent us from containing both of them. Plus, it introduce the prior of **X** into the covariance so that can functioning as a regularized term.

Question 15:

1. *Compute the objective function* $-\log(p(\mathbf{Y}|\mathbf{W})) = \mathcal{L}(\mathbf{W})$ *for the marginal distribution in Eq. 23.*
2. *Compute the gradients of the objective with respect to the parameters* $\frac{\partial \mathcal{L}}{\partial \mathbf{W}}$.

1. The equivalent of Type-II Maximum-Likelihood is shown in eq 42. The objective function would be similar to this:

$$\begin{aligned}
&-\log\left(p\left(\mathbf{Y}|\mathbf{W}\right)\right)\\
=&\mathcal{L}\left(\mathbf{W}\right)\\
=&-\log\left(\left(2\pi\right)^{-ND/2}|\Sigma_{\mathbf{Y}|\mathbf{W}}|^{-N/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{N}\mathbf{y}_i^T\left(\Sigma_{\mathbf{Y}|\mathbf{W}}\right)^{-1}\mathbf{y}_i\right)\right) \qquad (43)\\
=&\frac{ND}{2}\ln\left(2\pi\right)+\frac{N}{2}\ln\left(|\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I}|\right)+\frac{1}{2}\sum_{i=1}^{N}\mathbf{y}_i^T\left(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I}\right)^{-1}\mathbf{y}_i.
\end{aligned}$$

2. We have such differential identities:

$$\partial\,\mathrm{Tr}\left(\mathbf{X}\right)=\mathrm{Tr}\left(\partial\mathbf{X}\right) \qquad (44)$$

$$\partial\ln\left(|\mathbf{X}|\right)=\mathrm{Tr}\left(\mathbf{X}^{-1}\partial\mathbf{X}\right) \qquad (45)$$

$$\partial\mathbf{X}^{-1}=-\mathbf{X}^{-1}\left(\partial\mathbf{X}\right)\mathbf{X}^{-1} \qquad (46)$$

Where Tr($\mathbf{X}$) means $\sum_{i=1}^{N}x_{i,i}$. Then we can first re-write eq 43 into:

$$\mathcal{L}(\mathbf{W})=\frac{ND}{2}\ln\left(2\pi\right)+\frac{N}{2}\ln\left(|\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I}|\right)+\frac{1}{2}\mathrm{Tr}\left(\mathbf{Y}^T\left(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I}\right)^{-1}\mathbf{Y}\right). \qquad (47)$$

Then, we can get the gradients as:

$$\frac{\partial\mathcal{L}}{\partial\mathbf{W}}=N\left(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I}\right)^{-1}\mathbf{W}-\left(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I}\right)^{-1}\mathbf{Y}\mathbf{Y}^T\left(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I}\right)^{-1}\mathbf{W}. \qquad (48)$$

2.2 Practical

2.2.1 Linear Representation Learning

Question 16:

1. *Plot the representation that you have learned (hint: plot* $\mathbf{X}$ *as a two-dimensional representation).*

2. *Explain the outcome and discuss key features, elaborate on any invariance you observe. Did you expect this result?*

3. *How is the effect of representation learning dependent on the number of available samples? Please test lower values of N and discuss the observed implications.*

In this problem, the $\mathbf{W}$ is to be estimated through a Type-II Maximum-Likelihood, which is equivalent to minimize $\mathcal{L}(\mathbf{W})=\frac{N}{2}\ln\left(|\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I}|\right)+\frac{1}{2}\mathrm{Tr}\left(\mathbf{Y}^T\left(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I}\right)^{-1}\mathbf{Y}\right)$. Also, since we know that $\mathbf{Y}=\mathbf{X}^{'}\mathbf{W}^T$, we can get the learned $\mathbf{X}^{'}$ by using such equation:

$$\hat{\mathbf{X}}'=\mathbf{Y}\hat{\mathbf{W}}\left(\hat{\mathbf{W}}^T\hat{\mathbf{W}}\right)^{-1}. \qquad (49)$$

1. The learned representation is illustrated in fig 18.

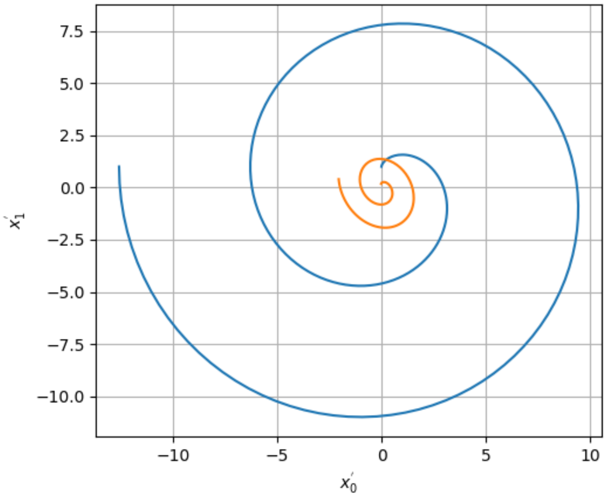


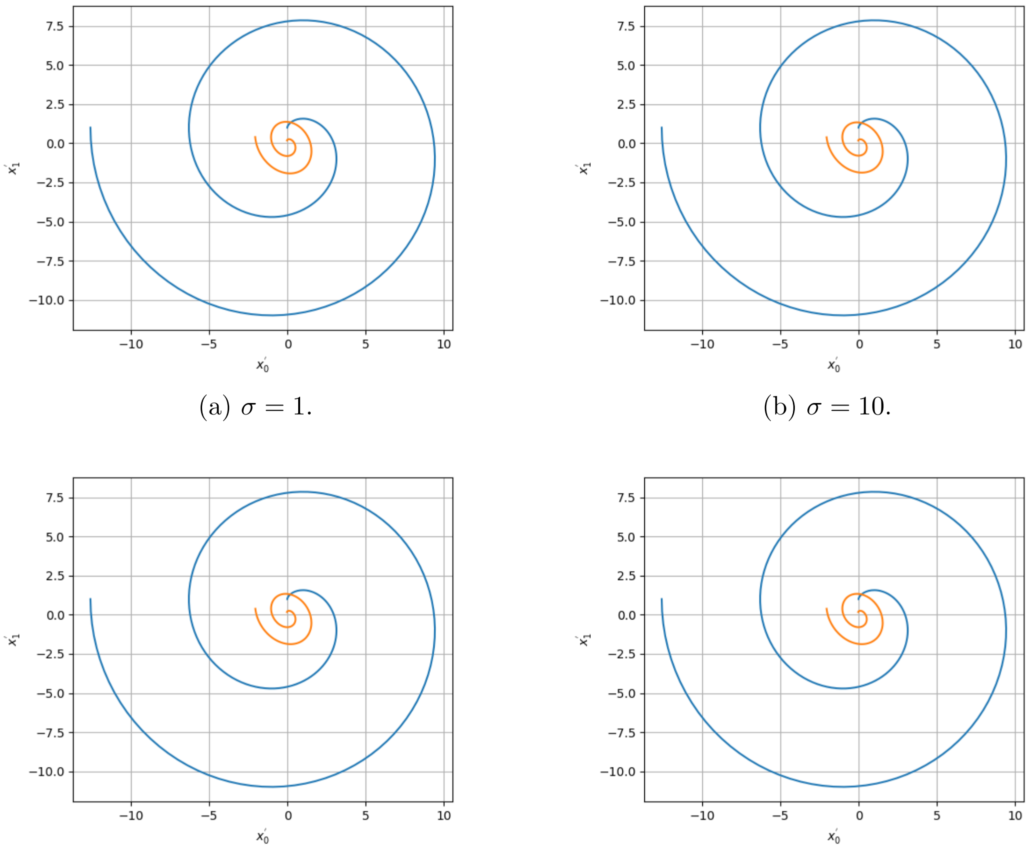
Figure 18: Learned $\mathbf{X}'$ (yellow curve) and ground truth (blue curve).

2. Overall, the result is expected. From fig 18, we can observe the difference between the ground truth and the representation learned. Two curves have similar shape, but the learned $\mathbf{X}'$ is scaled and rotated version of the ground truth. The likelihood distribution of $\mathbf{Y}$ given $\mathbf{W}$ is shown in eq 39, and we assume a rotation matrix $\mathbf{R}$ so that $\hat{\mathbf{W}} = \mathbf{WR}$, then we would have

$$\begin{aligned} P(\mathbf{Y}|\hat{\mathbf{W}}) &= \prod_{i=1}^{N} \mathcal{N}\left(\mathbf{y}_i|\mathbf{0}, \mathbf{WRR}^T\mathbf{W} + \sigma^2\mathbf{I}\right) \\ &= \prod_{i=1}^{N} \mathcal{N}\left(\mathbf{y}_i|\mathbf{0}, \mathbf{WW}^T + \sigma^2\mathbf{I}\right), \end{aligned} \tag{50}$$

   meaning the learned representation is invariant to the rotation of $\mathbf{W}$. Also, after changing the value of σ, which is the diagonal value of the covariance in the prior distribution of $\mathbf{X}$, same learned result is given (to make it reasonable, same random seed is used, and the result is shown in fig 19). The reason why the covariance in the prior distribution of $\mathbf{X}$ is invariant is because when we use the $\hat{\mathbf{W}}$ to recover the learned data, the covariance is offset.

3. The effect of representation learning does not dependent on the number of available samples too much. From fig 20, we can see that with different number of available samples, the learned representation are still in the same pattern with same rotation and scale. That's because when we are using the Type-II Maximum-Likelihood to estimate, we marginalize the observed $\mathbf{X}'$ so that the average effect of $\mathbf{X}'$ is taken into consideration. Then as long as the number of samples is enough to cover 2π, it is enough to recover the representation.



(a) $\sigma = 1$.

(b) $\sigma = 10$.

(c) $\sigma = 100$.

(d) $\sigma = 1000$.

Figure 19: Learned representation with different σ values.

3 The Evidence $p(\mathcal{D})$

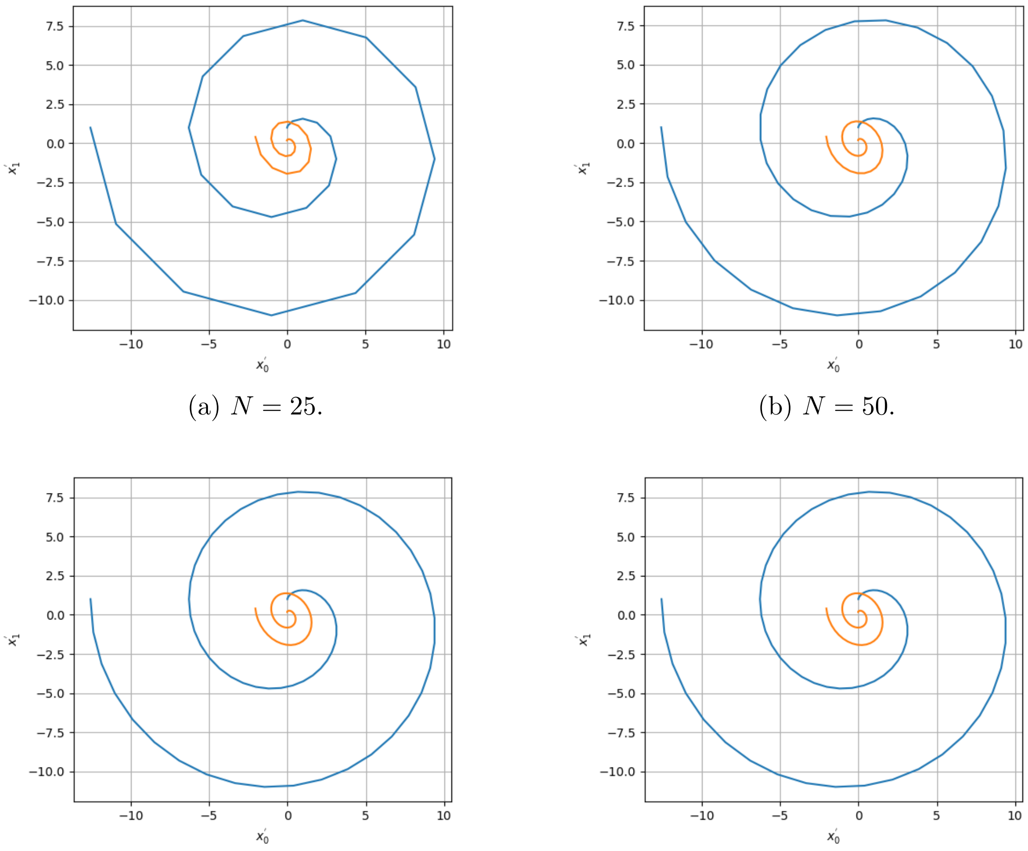
3.1 Theory

3.1.1 Data

3.1.2 Models

Question 17: *Why is this the simplest model, and what does it actually imply? What makes it a bad model on the one hand, and a good model on the other hand?*

As said, "the simplest model is something that simply takes all its probability mass and places it uniformly over the whole data sapce", so this is the simplest model because it's using the discrete uniform distribution for all the probability mass which is independent to the parameters. It implies that the probability for each data set is the same condition on the model and the model parameters, which is $\frac{1}{152}$. The advantage of such model is that, for a large range of data, this model has the largest evidence when compared to other models in [2]. However, on the other hand, it lose the ability to give high evidence when the data complexity is low. Also, since it's not using any parameters, the flexibility is bad.



(a) $N = 25$.

(b) $N = 50$.

(c) $N = 75$.

(d) $N = 100$.

Figure 20: Learned representation with different number of available samples.

Question 18: *Explain how each separate model works. In what way is this model more or less flexible compared to M_0? How does this model spread its probability mass over $\mathcal{D}$?*

For the model M_1, the likelihood of the data is

$$p(\mathcal{D}|M_1, \boldsymbol{\theta}_1) = \prod_{n=1}^{9} \frac{1}{1 + e^{-t^n \theta_1^1 x_1^n}}. \tag{51}$$

Then each separate model would be

$$p(t^i|M_1, \theta_1^1) = \frac{1}{1 + e^{-t^i \theta_1^1 x_1^i}}, \tag{52}$$

where $i \in \{1, \cdots, 9\}$. Each separate model is a logistic regression without bias and ignores the second dimension of $\mathbf{x}$. For each dimension of the target, the probability of having t^i (which is binary and $t^i \in \{-1, 1\}$) follows a modified logistic regression only based on the x_1 dimension and the corresponding parameter.

It is more flexible than M_0, because it take the location (only dimension x_1) and the corresponding weight (θ_1^1). However, M_1 focus on the data sets that separate in the dimension of x_1 which lower it's flexibility to the excluding data sets.

Since the model only use x_1, it would give high probability to the data set like the data set (b) in fig 21. And when doing the classification, this model would only classify based on this dimension. Besides, since no bias used in the exponent of the likelihood, the data set with boundary at $x_1 = 0$ is preferred and have higher probability.

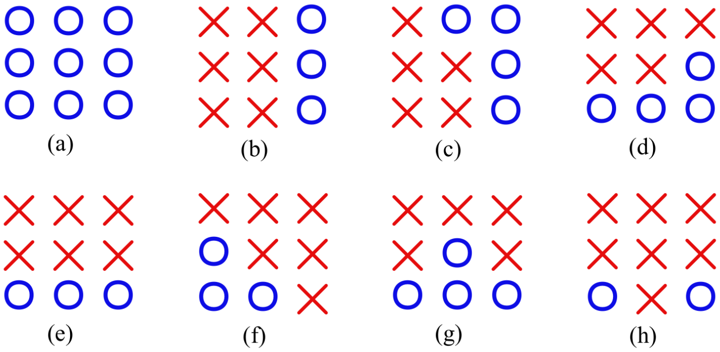


Figure 21: Different data sets, where red cross means -1 and blue circle means $+1$ [2].

Question 19: *Discuss and compare the models. In particular, please address the following questions in your discussion?*

- *How have the choices we made above restricted the distribution of the model?*
- *What datasets is each model suited to model? What does this actually imply in terms of uncertainty?*
- *In what way are the different models more flexible and in what way are they more restrictive?*

- M_0 treats all data sets equally; M_3 is standard logistic regression; M_2 is the same as M_3 but without the bias weight θ_3^3; M_1 is the same as M_2 except it ignores the second dimension of $\mathbf{x}$ [2]. Meaning M_0 would give equal to all the data sets. M_1 would give higher probability to data sets with boundary vertically crossing the origin. M_2 would give higher probability to data sets with boundary crossing the origin. M_3 would allow the data set have boundary with offset.

- For M_0, all the data sets are treated equally, so no data set is more suitable than others in the case of M_0. For M_1, only data sets with boundary vertically crossing the origin would be suitable. For M_2, data sets with boundary crossing the origin would be suitable. For M_3, data sets with boundary that have offset from the origin would be suitable. In the terms of uncertainty, if the data sets are not "suitable" for a model, the model would have low evidence and high uncertainty.

- M_3 is a standard logistic regression that take bias and both dimensions into consideration, thus can be considered as the most complex and flexible model. Unlike, M_2 is restricted to the origin and M_1 is restricted to the dimension, M_3 takes all the parameters and can realize the same functions as M_1 and M_2 by setting the parameter to zero. However, this means it have bigger range in data sets that can have high probability and when the data is actually simple enough to use M_1 and M_2, the use of M_3 would give more uncertainty.

3.1.3 Evidence

Question 20: *Explain the process of marginalisation and briefly discuss its implications in the given context of the model evidence.*

To obtain the model evidence $p(\mathcal{D}|M_i)$, the likelihood over $\mathcal{D}$ is marginalized over all possible values of the parameters by specifying a prior over $\boldsymbol{\theta}$, so that

$$p(\mathcal{D}|M_i) = \int_{\forall \boldsymbol{\theta}} p(\mathcal{D}|M_i, \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{53}$$

Here, the prior distribution of $\boldsymbol{\theta}$ can be viewed as the weight for different states of $\boldsymbol{\theta}$. In the given context of the model evidence, it means we take all the possible boundaries for each model into consideration to give an weighted average likelihood over the observed data set when given the model.

Question 21: *What does this choice of prior imply? How does the choice of the parameters of the prior $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ affect the model?*

The prior distribution is

$$p(\boldsymbol{\theta}|M_i) = \mathcal{N}\left(\boldsymbol{\theta}|\mathbf{0}, \sigma^2 \mathbf{I}\right), \tag{54}$$

where $\sigma^2 = 10^3$.

This choice of prior implies that we think that the different $\boldsymbol{\theta}$ in a wide range centered at origin have almost the same probability. Also, since it follows a multivariate normal distribution with zero mean and diagonal covariance where elements on the diagonal are the same, we think the $\boldsymbol{\theta}$ at the origin have the largest probability and each parameters are independent to each other. Since the range of parameters can be chosen from is large, meaning the model would be free to choose a various parameters' values and they are almost equally possible to each.

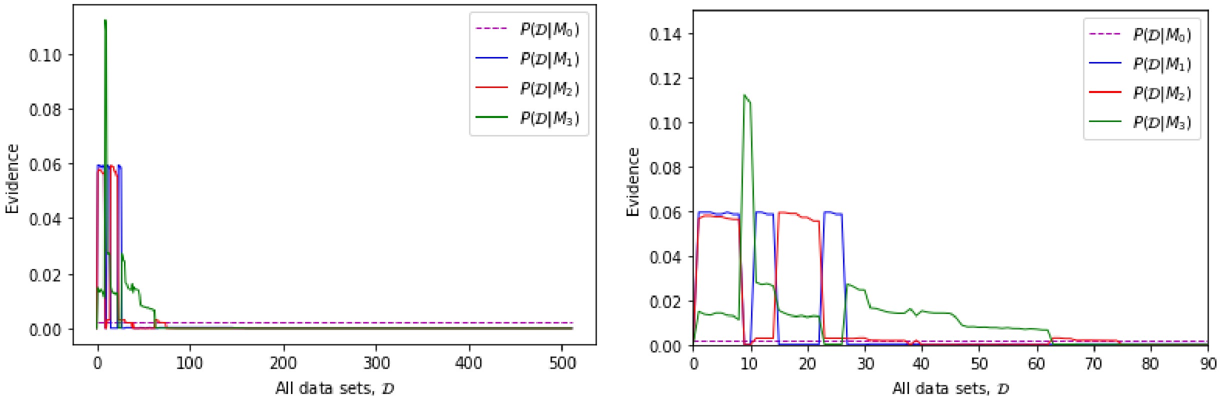
3.2 Practical

Question 22: *Plot the evidence over the whole dataset for each model (and sum the evidence for the whole of $\mathcal{D}$, explain the numbers you get). The **x-axis** index the different instances in $\mathcal{D}$ and each models evidence is on the **y-axis**. How do you interpret this? Relate this to the parametrisation of each model.*

The evidence over the whole data set for each model is obtained by a Monte Carlo approach with sampling on the prior of $\boldsymbol{\theta}$. The result is illustrated in fig 22. To sample, I used $S = 10^4$ rather than 10^8.

From the plot, we can see that M_0, which is dashed, is always a constant over the data sets which is expected from the likelihood. Model M_1, which is blue, have high probability where the data sets can be vertically separated (according to x_1) and such data is near the origin. Model M_2, which is red and have high evidence where the evidence of M_1 is high (this means M_2 can have the same ability of M_1 by setting the parameter of x_2 to zero). Besides, M_2 also spread high evidence to other data sets which have boundary pass original location. The last model M_3, which is green, spreads a larger range of data sets and cover the range of M_1 and M_2 at the same time, which indicated that M_3 is the most complicated and flexible and is the same from the conclusion above.

The sum of evidence for each model over the whole possible data sets is 1, which is because the evidence of a single data set means the likelihood of having this data set by using the model, thus it is reasonable that the sum over the whole data set would give 1.



(a) All data sets, $\mathcal{D}$.

(b) Subset of possible data sets, $\mathcal{D}$.

Figure 22: Evidence over the whole data sets for each model.

If we use another way to arrange the data sets by using the binary number, so that the digits follows the following matrix:

$$\begin{bmatrix} 8 & 7 & 6 \\ 5 & 4 & 3 \\ 2 & 1 & 0 \end{bmatrix}$$

we would get the evidence plot illustrated in fig 23.

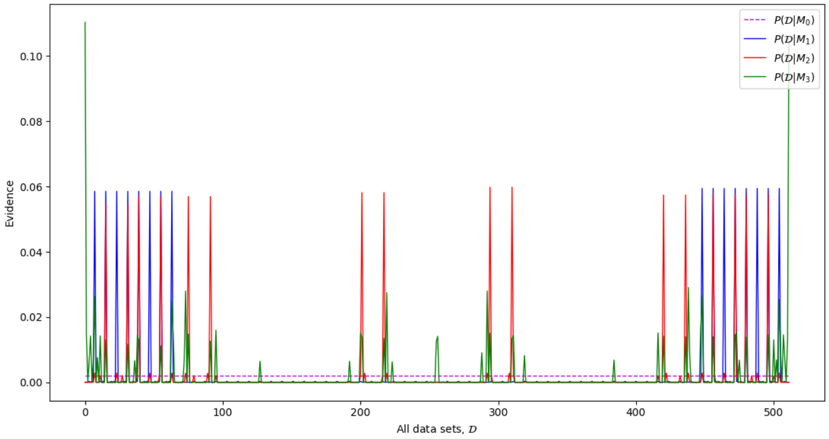
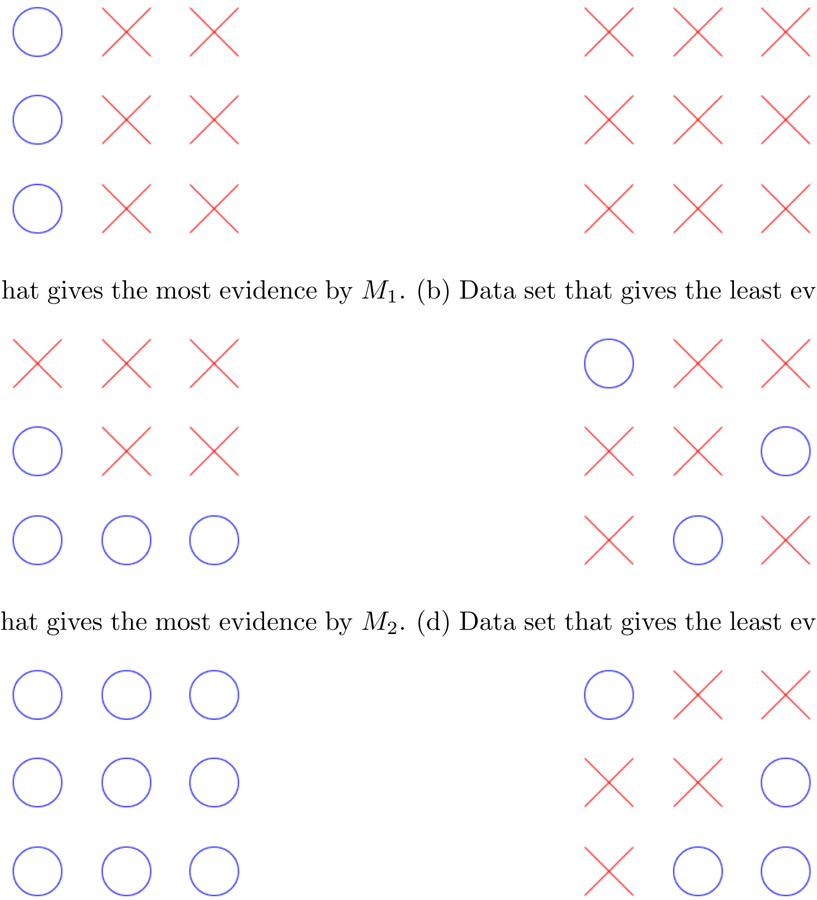


Figure 23: Evidence over the symmetrically arranged data sets for each model.

Question 23: *Find **using np.argmax and np.argmin** which part of the $\mathcal{D}$ that is given most and least probability mass by each model. Plot the datasets which are given the highest and lowest evidence for each model. Discuss these results, do the findings make sense?*

For M_0, since all the data sets are treated equally, it's not meaningful to plot the data set with lowest and highest evidence. For M_1, M_2, and M_3, the data sets that give the most and the least probability mass is illustrated in fig 24. The data implies the same idea we gave about which boundaries are preferred by each model.



(a) Data set that gives the most evidence by M_1. (b) Data set that gives the least evidence by M_1.

(c) Data set that gives the most evidence by M_2. (d) Data set that gives the least evidence by M_2.

(e) Data set that gives the most evidence by M_3. (f) Data set that gives the least evidence by M_3.

Figure 24: Data sets that give the most and least probability mass by each model.

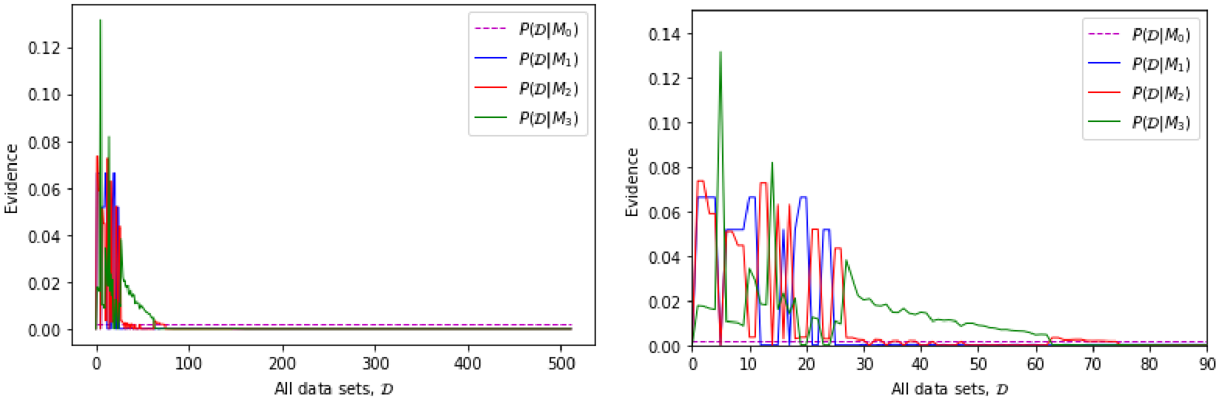
Since M_1 only depends on x_1, it would give the highest evidence to the data that have boundary vertically and pass origin of location, which is just fig 24a. And for fig 24b, since all the target values on all the locations are the same, which means there's no boundary, it is reasonable that it have the lowest evidence.

For M_2, since it depends on x_1 and x_2 but have no bias, the highest evidence must occur at the data set that have the boundary passing the origin of location, which is the same the fig 24c describes. For the data set shown in fig 24d, since it requires multiple boundaries to separate the data, it is also reasonable that M_2 would have the least evidence at this data set.

For M_3, since it is depends on all three parameters and is the most flexible, it can take the bias into consideration. The fig 24e shows that it can support biased data sets very well. But still, fig 24f, where the data set requires two boundaries to separate, shows that it have low evidence on such data sets.

Question 24: *What is the effect of the prior $p(\boldsymbol{\theta})$.*

- *What happens if we change its parameters?*
- *What happens if we use a non-diagonal covariance matrix for the prior?*
- *Alter the prior to have a non-zero mean, such that $\mu = [5, 5]^T$?*
- *Redo evidence plot for these and explain the changes compared to using zero-mean.*



(a) All data sets, $\mathcal{D}$.

(b) Subset of possible data sets, $\mathcal{D}$.

Figure 25: Evidence over the whole data sets for each model with non-zero mean in prior.

- If we change the parameters of the prior $p(\boldsymbol{\theta})$, meaning we change our preferred data sets. We would get a different distribution which mean the weight of different $\boldsymbol{\theta}$ would be changed. If the a smaller σ value is used, we would be more certain about the parameters around the mean and allow for a smaller range of possible values of $\boldsymbol{\theta}$. Meaning the models would give more sharper evidence on $\mathcal{D}$.

- If we use a non-diagonal covariance matrix for the prior, the parameters in $\boldsymbol{\theta}$ would be dependent to each other. Then if two parameters are positively correlated, we will be certain on more data sets.

- We use this non-zero mean such that we believe $\boldsymbol{\theta}$ is more likely to be this mean value. The evidence over the whole data sets for each model using a non-zero mean in the prior, such that $\mu = [5, 5]^T$ is illustrated in fig 25. From the plot, we can see that there are more sharp spikes in a smaller range of data sets when compared to a non-zero mean situation. Also, if we arrange the data sets in a symmetric way, we can see that the evidence is biased to the left of plot.

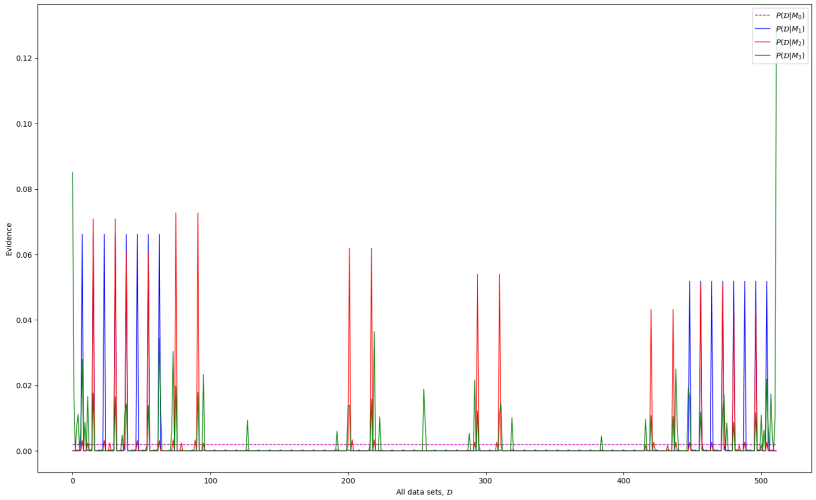


Figure 26: Caption

References

[1] C. M. Bishop, *Pattern recognition and machine learning.* Information science and statistics, New York: Springer, 2006.

[2] I. M. Z. Ghahramani, “A note on the evidence and bayesian occam’s razor,” tech. rep., Gatsby Unit Technical Report GCNU-TR 2005–003, 2005.